بسم الله الرحمن الرحيم

Biostatistics

L IV

15th –July 2024

PROF. DR. WAQAR AL-KUBAISY

Description statistics summarization

Presentation — **Numerical**

Graph — **Table**

-this approach might not be enough,

-comparisons between one set of data & another

-summarize data by one more step further .

-presenting a set of data by a

- single Numerical value

The central value as representative value in a set of data,

1-Measures of central tendencies (Location) .

A value around which the data has a tendency to congregate (come together)or cluster

2-Measures of Dispersion, scatter around average

A value which measures the degree to which the data are or are not, spread out

1-Measures of central tendencies (Location) .
A value around which the data has a tendency to congregate (come together)or cluster
2-Measures of Dispersion, scatter around average
A value which measures the degree to which the data are or are not , spread out

The central value as

1-Measures of central tendencies (Location)

75, 75, 75, 75, 75, 75, Mean = ????

75, 70, 75. 80, 85. Mean = ????

60, 65, 55, 70, 75, 75, ,70, 80, Mean= ????

$$\overline{X} = \frac{\sum X}{N}$$

2-Measures of Dispersion,

The central value as
1-Measures of central tendencies
2-Measures of Dispersion,

Measures of Dispersion
(Measures of Variation)
(Measures of Scattering)
Measures of spread

Measures of Dispersion
(Measures of Variation)
(Measures of Scattering)
measures of spread

Measures of Dispersion

1- **Range**

2-Interquartile range

3- Variance

4- Stander Deviation

5- Coefficient of variance

the choice of the most appropriate measure depends crucially on the type of data involved

Measures of spread

Measuring of spread are very useful.

There are three main measures in common use .

once again the type of data influence the choice of an appropriate measure

the choice of the most appropriate measure depends crucially on the type of data involved

1- Range
2-Interquartile range
3- Variance
4- Stander Deviation
5- Coefficient of variance

The Range

simplest

most obvious one of dispersion.

It is the distance from the smallest to the largest

It *Obtained by*

subtracting lowest value from the highest value in a set of data .

Pulse rate 70 76 74 78 72 74 76

Range = 78 – 70 =

The range is best written
like rang of data (from- to) 70-78
rather than single-valued difference which is much less informative

- **The range is not affected by skewness**

70 72 74 76 76 78 78 **78-70** **70-78**

sensitive to the addition or removal of an outlier value

66 70 74 90, 100 120 124 **124-66** **66-124**

66 70 74 90, 100 120 124 545 **66-545**

Its disadvantage

it is based on **only two observations**

(the lowest and highest value) and

- give no idea about others,
- not take into consideration other values in data
- sensitive to an **outlier value** **Therefore**
- **It is not very useful measures of variation,**
- ✓ **because it does not use other observation**

Therefore ;

Therefore ;

Sensitive an outlier value

Interquartile rang (I q r).

- ✓ ***measure the variation of one observation from the other***
- ✓ **Standard deviation**

Standard deviation

75, 70, 75. 80, 85.
60, 65, 55, 70, 75, 75, ,70, 80,

variation of each value, from the other??

60, 65, 55, 70, 75, 75, ,70, 80, 40, 45, 53, 77, 75, 95, ,100, 88, 68, 95, 57, 78, 35, 95, ,78, 85, 67, 69, 35, 71, 79, 77

variation of each value from the other???

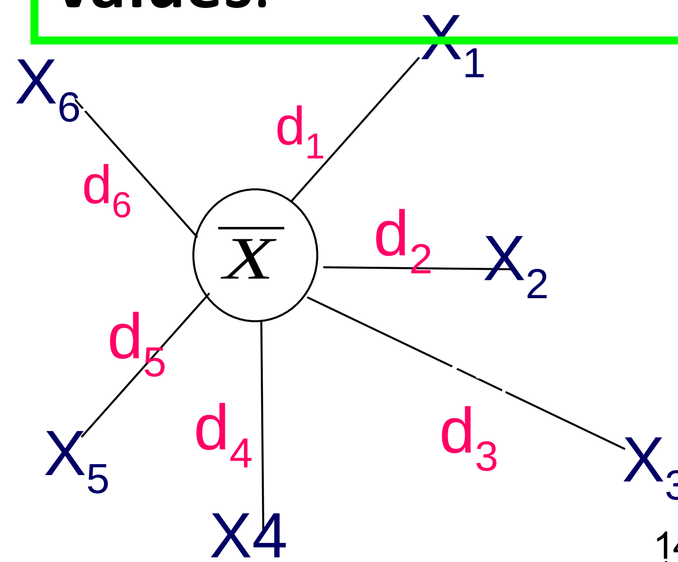
13

75, 70, 75. 80, 85. **Mean = ????**

60, 65, 55, 70, 75, 75, ,70, 80, Mean= ????

$$\overline{X} = \frac{\sum X}{N}$$

the mean (average) variation of all data values from the over all mean of all values.

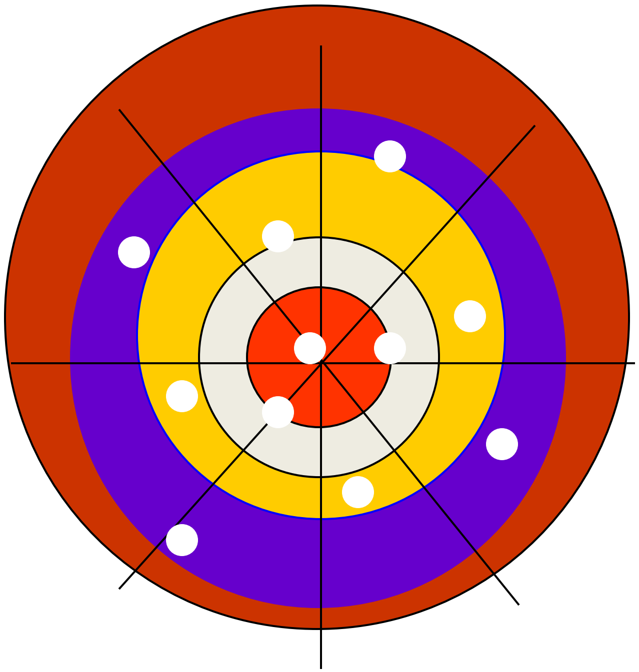


the mean (average) distance of all data values from the over all mean of all values.

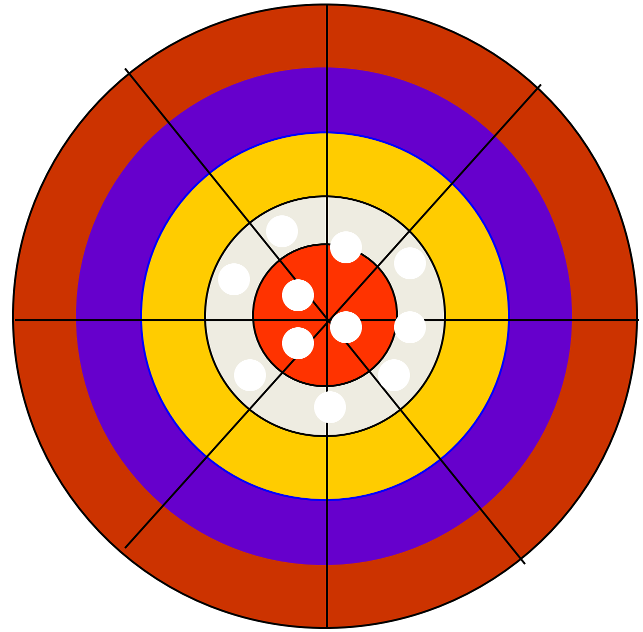
14

Measures of Dispersion

SHOOTER A



SHOOTER B



Both shooters are hitting around the "centre" but shooter B is more "accurate"

- **The smaller the mean distance is**

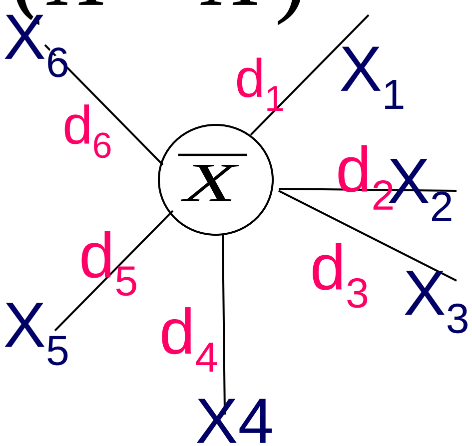
✓ **the narrower the spread of values**

Measures of Dispersion

student No.	score	$x - \overline{x}$
1^{st}	6	$6 - 3 = +3$
2^{nd}	2	$2 - 3 = -1$
3^{rd}	4	$4 - 3 = +1$
4^{th}	1	$1 - 3 = -2$
5^{th}	3	$3 - 3 = 0$
6^{th}	2	$2 - 3 = -1$

the **mean**(average) variation of **all** data **values** from the **over all mean** of all values.

$$(X - \overline{X})$$



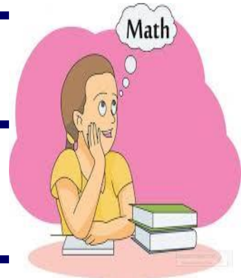
$$\sum X = 18$$

$$\overline{X} = 3$$

$$\sum (X - \overline{X}) = zero$$

????

student No.	Score	$X - \overline{X}$	$(X - \overline{X})^2$
1st	6	6 – 3 = +3	9
2nd	2	2 – 3 = -1	1
3rd	4	4 – 3 = +1	1
4th	1	1 – 3 = -2	4
5th	3	3 – 3 = 0	0
6th	2	2 – 3 = -1	1



$$\overline{X} = 3$$

$$\sum X = 18$$

$$\sum(X - \overline{X}) = zero$$

$$\sum(X - \overline{X})^2 = 16$$

$$(X - \overline{X}) \rightarrow (X - \overline{X})^2$$

Variance

$$S^2 = \frac{\sum(X - \overline{X})^2}{N - 1} =$$

$$S^2 = \frac{16}{5}$$

3.2 $score^2$

????

Variance S^2

It is the Average of squared deviation of observation from the mean in a set of data .

$$S^2 = \frac{\sum (X - \overline{X})^2}{N - 1}$$

3.2 score^2

????

The Disadvantage or drawback of variance

that its unit is squared Kg^2 , bacteria^2 ….., So Restore the squared unit into its **original form**

by

taking the square root of this (S^2) value, this is **known as Stander Deviation (S.D).**

$\sqrt{3.2}$ = ±1. 789 score

Standard Deviation ± S.D.

It is the square root of variance. S.D = √S²

$$S^2 = \frac{\sum (X - \overline{X})^2}{N - 1}$$

$$\sqrt{\frac{\sum (X - \overline{X})^2}{N - 1}} = \pm S.D$$

± S.D (S) it is the square root of the **Average square deviation** of **observation from the mean in a set of data**

One advantage of SD is that unlike the iqr
it uses all the information in the data

Steps in calculating S.D

1.Determine the **mean** $\overline{X}$

2-Determine the deviation of each value from the mean $(X - \overline{X})$

3-.**Square each deviation of value from mean** $(X - \overline{X})^2$

4-Sum these square deviation of value from mean $\sum(X - \overline{X})^2$

(sum of square) . $\sum(X - \overline{X})^2$

5-**Divide** this **square deviation of value from mean by N-1**

$$\frac{\sum(X - \overline{X})^2}{N-1}$$

.

6-Take the square root of deviation of value from mean by N-1

$$\sqrt{\frac{\sum(X - \overline{X})^2}{N-1}} = \pm S.D$$

Short Cut Method

	score	Score ²
1	6	36
2	2	4
3	4	16
4	1	1
5	3	9
6	2	4
total	**18**	70

$$S^2 = \frac{\sum (X - \overline{X})^2}{N - 1}$$

$$\sum (X - \overline{X})^2 = \sum X^2 - \frac{(\sum X)^2}{N}$$

$$S^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N - 1}$$

$$\frac{70 - 18 \times 18/6}{5} = \quad 70-54=16/5= 3.2$$

$\sqrt{3.2}$ = 1.789???????

Short Cut Method for S.D

1-Square **each absolute** individual value . X^2

2-Sum these **squared values** $(\sum X)^2$.

3-Sum the all absolute value of observation $X_1 + X_2 + X_3 + \ldots = \sum X$

4-Square this **sum** of absolute values

5-Divide this sum of absolute values by N $\frac{(\sum X)^2}{N}$

6-Subtract $\frac{(\sum X)^2}{N}$ from $\sum X^2$ $\longrightarrow$ $\sum X^2 - \frac{(\sum X)^2}{N}$ (**sum of square**)

7-Divided all this result by N-1 , $S^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N-1}$

8-Take the square root of this last result,

$$S.D = \pm\sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N-1}}$$

Example

Short Cut Method

	Score	Freq.(No.of Students)	XF	X^2F
	6	2	$6\times2=12$	$6^2\times2=72$
	2	4	$2\times4=8$	$2^2\times4=16$
	4	3	$4\times3=12$	$4^2\times3=48$
	1	5	$1\times5=5$	$1^2\times5=5$
	3	2	$3\times2=6$	$3^2\times2=18$
	2	6	$2\times6=12$	$2^2\times6=24$
total		22	55	183

$$S^2 = \frac{\sum (X-\overline{X})^2}{N-1}$$

$$\sum (X-\overline{X})^2 = \sum X^2 - \frac{(\sum X)^2}{N}$$

$$S^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N-1}$$

$$S^2 = \frac{183 - \frac{55^2}{22}}{22-1} = \frac{183-137.5}{21} = 2.166$$ scor²

S.D = √2.166 = 1.472

??????

Disadvantage Limitation or Drawback of S.D

It is depend on the unit of measurement,

we can't compare between two or more data

to overcome this

Coefficient of Variation C.V

It is representing by measuring the variation in relation to the percentage of mean of that data

$$C.V = \frac{S.D}{\overline{X}} \times 100$$

$$C.V = \frac{1.47}{2.5} X100 = 58.8\%$$

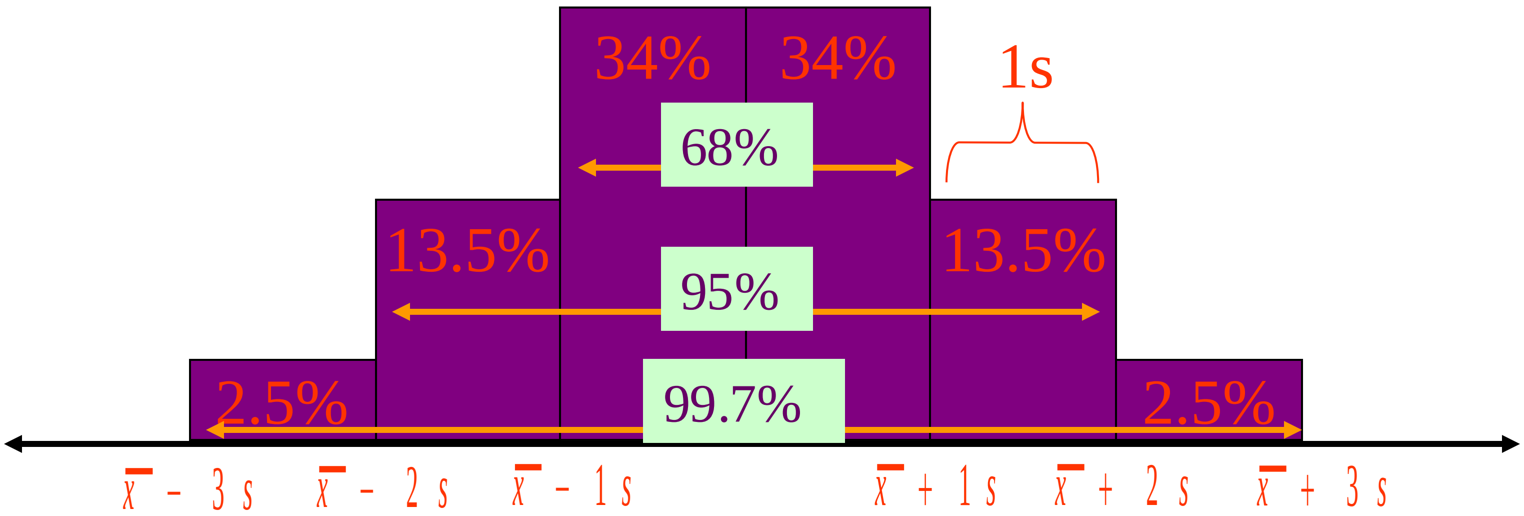
????

-C.V is used

to compare between two or more data

- with different units of measurement .
- data with large difference between their means .

Interpreting Standard Deviation



For bell-shaped shaped distributions, the following statements hold:

- Approximately 68% of the data fall between $\bar{x} - 1s$ and $\bar{x} + 1s$
- Approximately 95% of the data fall between $\bar{x} - 2s$ and $\bar{x} + 2s$
- Approximately 99.7% of the data fall between $\bar{x} - 3s$ and $\bar{x} + 3s$

For NORMAL distributions, the word 'approximately' may be removed from The above statements.

Q1

SD used with median

SD used with rang

SD used in nominal data

IQR used with the mean

Variance is the best measurement of dispersion

Q2 Measures of dispersion are

1

2

3

4

5

6

Thank You

1. Median is the value with a highest frequency
2. When the data is skewed , median is the appropriate measures of CT
3. Mean is appropriate measures of Ct in ordinal data

4. Mode used when we have Metric continuous data

5- mean is unique what ever the size of data is

Q1

Thirty (30) pregnant women attending Al- Karak antenatal clinic during 23-februry 2023 showing gain in weight as follows:

Weight gain (kg	**NO.of women**
4	3
7	5
10	10
12	8
16	4

1-Present this data graphically,

2- Compute the measures of Central tendency

3- Compute Measures of Dispersion

Therefore ;

Sensitive an outlier value

Interquartile rang (I q r).

- ***measure the variation of one observation from the other***
- **Standard deviation**

Percentile

A percentile provides information about how the data are spread over the interval from the smallest value to the largest value.

The pth percentile (25%) (30%):
is a value such that at least p percent of the observations are less than or equal to this value **and**
at least (100 - p) (75%) (70%) percent of the observations **are greater than or equal to this value**.

The pth percentile is a value so that **roughly p% of the data are smaller and (100-p)% of the data are larger**.

Three Steps

Three Steps for computing a percentile.

1. **Sort the data from low to high;**
2. **Count the number of values (n);**
3. **Select the p*(n+1) observation**

Examples

The following data represents cotinine levels in saliva (ng/ml) after smoking. We want to compute the 50th percentile.

73, 58, 67, 93, 33, 18, 147

Sorted data: 18, 33, 58, 67, 73, 93, 147

There are n=7 observations.

Select 0.50*(7+1)=4th observation.

Therefore, the **50th percentile** equals 67.

Notice that there are

three observations larger than 67 and

three observations smaller than 67.

Examples

The following data represents cotinine levels in saliva (ng/ml) after smoking. We want to compute the 20th percentile.

73, 58, 67, 93, 33, 18, 147

Sorted data: 18, 33, 58, 67, 73, 93, 147

Suppose we want to compute the 20th percentile.

Notice that $p*(n+1)$ = **$0.20*(7+1)=1.6$.** This is not a whole number so we select halfway **between 1st and 2nd** observation

they have to go **six tenths of the way to the** second value.

Calculation of percentile value

The pth percentile is
the value in the p/100 (n+1) th position.

For example
the 20th percentile

Calculation of percentile value

the birth weight(grm) of 30 infants which we put in ascending order.

2860	2994	3193	3266	3287	3303	3388
3399	3400	3421	3447	3508	3541	3594
3613	3615	3650	3666	3710	3798	
3800	3886	3896	4006	4010	4090	4094
4200	4206	4490				

Calculation of percentile value

> **The pth percentile is**
> **the value in the p/100 (n+1) th position.**

the 20th percentile is the

20/100(n+1)   with the BW values

20/100 (30 +1)

0.2x31 observations= 6.2observation

the birth weight of 30 infants which we put in ascending order.

2860 2994 3193 3266 3287 3303 3388 3399 3400
3421 3447 3508 3541 3594 3613 3615 3650 3666
3710 3798 3800 3886 3896 4006 4010 4090 4094
4200 4206 4490

Cont. ..Calculation of percentile value

The 6th value is 3303 g

the 7th value is 3388 g

a difference of 85 g

the 20th percentile is

3303 + 0.2 of 85 g

which is

3303g + 0.2x 85 g =

=3303g+17g

= 3320 g

the birth weight of 30 infants which we put in ascending order.

2860	2994	3193	3266	3287	
3303	3388	3399	3400	3421	3447
3508	3541	3594	3613	3615	3650
3666	3710	3798	3800	3886	
3896	4006	4010	4090	4094	
4200	4206	4490			

The pth percentile is

the value in the p/100 (n+1) th position.

Similarly we could calculate

the birth weight of 30 infants which we put in ascending order.

2860	2994	3193	3266	3287	3303
3388	3399	3400	3421	3447	3508
3541	3594	3613	3615	3650	3666
3710	3798	3800	3886	3896	4006
4010	4090	4094	4200	4206	4490

cont. ……Calculation of percentile value

the deciles

which subdivide the data values into 10 (not 100)equal division,

and

Quintiles

which sub-divide the values into five equal –sized groups

Collectively we call

- **percentiles,**
- **deciles** divide the sorted data into ten equal parts, so that each part represents 1/10 of the sample or population. **and**
- **quintiles**

The pth percentile is the value in the p/100 (n+1) th position.

A quartile is :

a division of observations into four defined 25% 50%

Interquartile rang (i q r).

One solution to the problem of the sensitivity to extreme value (outlier) is to

- ✓ **chop the quarter(25 percent) of the values of both ends of the distribution**
  (which removes any troublesome outliers)

then measure the range of the remaining values

- ❑ **this distance is called**
- ❑ **interquartile range or i q r .**

Calculation of iqr

To calculate iqr we need to determine two values

first quarantile (Q1)	third quarantile (Q3)
The value which **cuts off the bottom 25 percent of values**	**The value which cuts off the top 25 percent of values,**

The interquartile range is then written as (Q1 to Q3)

31X 0.25 = 7.75

31X .75 = 23.25

the birth weight of 30 infants which we put in ascending order.

2860	2994	3193	3266	3287	3303	3388	3399
3400	3421	3447	3508	3541	3594	3613	3615
3650	3666	3710	3798	3800	3886	3896	4006
4010	4090	4094	4200	4206	4490		

The pth percentile is
the value in the p/100 (n+1) th position.

with the BW data
Q1= 3396.25g and
Q3 = 3923.50 g

7.75th 3399-3388=11x.75=8.25+3388=
3396.25
0.75x 31=**23.25th**
4006-3896=110x.25=27.5+3896=3923.5

the birth weight of 30 infants which we put in ascending order.
2860 2994 3193 3266 3287 3303 3388 3399 3400
3421 3447 3508 3541 3594 3613 3615 3650 3666
3710 3798 3800 3886 3896 4006 4010 4090 4094
4200 4206 4490

Therefore iqr = 3369. 25 to 3923.50)g

the middle 50 percent

Calculation of iqr

the middle 50 percent of infant weighed between 3396.25 and 3923.50 g

✓ **The interquartile range**

indicate

- **the spread of the middle 50%of the distribution,**
- **together with the median is useful adjunct (accessory) to the range**
- **it is less sensitive to the size of the sample providing that this is not too small**

The interquartile range is not affected either by

Outlier

skewness

BUT

it does not use all of the information in the data set since it ignores the bottom and top quarter of values.

✓ ***measure the variation of one observation from the other***

✓ **Standard deviation**


Thank you!